GIS Data Naming Convention

The goal of the Data Standards and Recommendations Committee is to provide recommendations and guidelines to Indiana GIS user communities to facilitate the collection, maintenance and analysis of GIS data; and, to communicate existing federal, state and local data standards. The Data Standards and Recommendation Committee will not recommend software, hardware or operating systems. Furthermore, the Data Standards and Recommendation Committee will not impose any of these recommendations and guidelines as a requirement on any GIS user community
.

Why a naming convention?

GIS workers are often faced with selecting the one most appropriate dataset for a particular task from many possible choices. These choices become more complex when several sources of the same or similar data, from different jurisdictions are involved. This has been a problem for casual users trying to select data from various libraries on the web. This problem has been especially severe in situations when many datasets must be quickly brought together and synthesized on the fly for timely responses.

The underlying principle of this recommendation is to get to some level of consistency so that sharing of data becomes easier for the end user. Attainment of that principle is left to the steward of the data. It is understood that there may exist filename length constraints and other considerations imposed within certain software and hosting environments. In those instances where full implementation of this naming convention is not possible, it is left to the stewards of the data to reconcile how implementation of this recommendation can proceed within their respective organizations.

Who will find the file naming convention useful and relevant?

Consumers of GIS data working with data from cross or multiple jurisdictions will find this convention most useful, since it allows identification of several data characteristics from the name of the dataset. Creators of data will also find this naming convention useful if they use the convention to meet internal needs for consistency in their respective organizations. The time-

saving benefits of consistency in naming data have already been demonstrated. Those benefits are growing drastically with increasing use of GIS for emergency planning and mitigation.

The primary objective of this recommended best practice is to uniquely identify each dataset if several versions of one particular dataset are encountered. The secondary objective of this recommended best practice is to develop consistency in naming files across organizations so that users of data have a predictable way to recognize relevant data. Thirdly, and indirectly, we are striving to denote contacts should any questions arise regarding the data. Put another way, the recommendation seeks to denote what is in the dataset, the steward of the dataset, and optionally, the geographic extent of the data and either the date of the content or the date the file was created. It is believed that uniqueness of the filenames will also occur as a result of implementing the convention as recommended.

Recommendations

The Indiana Geographic Information Council recommends that GIS data be named using the following structure:

keyword_steward_extent_date

Explanation

Keyword – essential

The steward of the data assigns the keyword. The intent of the keyword is to be as descriptive of the contents of the data as possible by using a word or short phrase. A lookup table or lexicon of suitable keywords is not available at this time, but may be added in time. Assignment of keywords by the steward is recommended to assure consistency when several similar datasets are being made available. An example of when this may be necessary is if the steward of roads elected to make several discrete types of roads available like interstate highways, state roads, county roads and local roads, along with an all inclusive "roads" dataset. In this instance, it would be useful to have the steward determine how best to reflect the unique content of each "road" dataset with a keyword of their choice.

Steward – essential

The steward of a dataset is usually either the creator of the dataset or the last one to make a significant modification to a dataset. Stewardship is intended to reflect the entity that is the source of the data. In most instances this entity will be the creator of the data or the one that has engaged a contractor or third party to gather and develop the data. Other examples are forthcoming. However, for state government, the stewardship role should first be resolved to the agency. How the agency is denoted is left to the steward, or in this case the agency. It is not important what the agency is called (IDNR vs. DNR) rather, it is more important that an acronym or short name is selected and then used consistently for the duration.

In some instances, a dataset may be significantly modified by an entity other than the one that created the geometry. The group that made the modifications becomes the steward that is reflected in this part of the filename. For instance one group created a file of county boundaries and a second group has affixed county contacts for a specific program to the county boundary file. The county boundary file has now become a county contact file created by the second group. The second group is the steward of the contact information and should be denoted in the filename.

In other instances it may be necessary to consider compound steward names. The compound names can be used in different ways. Within state government, the first use may be to indicate a particular program or office within an agency (i.e. IDEMland or IDEMwater to indicate IDEM Office of Land Quality or IDEM Office of Water Quality respectively). A second use may be to indicate the provenance of the data from a source other than the agency currently hosting the data. For instance, should Department of Natural Resources host aerial photography gathered by the United States Department of Agriculture – an acceptable solution would be DNRusda. A third use of a compound steward name would be when agencies share collection and maintenance of a dataset, then showing both agency names would be acceptable too. Use of compound names is up to the host and partner agencies, keeping in mind overall filename length, simplicity and consistency for the duration that the data are to be shared.

Extent – optional

The geographic extent may be included as a clue to the resolution of the data, international versus county, or to indicate that the data are truly of the extent indicated. Recommended modifiers indicating extent are as follows:

> _IN = statewide extent
> _US = national extent
> _INT = international extent (i.e. Great Lakes, parts of Canada)
> _REG = regional (multi-county)
> _CO = single county
> _CI = city or municipality

Date – optional

The date may be used two ways. First, the date may be used to indicate the date of the content. Depiction of the boundaries of a municipality at a given time in the past would be an example of the date of the content. In cases where a dataset is updated periodically it may be beneficial to indicate the date that the file was created instead of the age range of the content.
The recommended format for indicating a date is:
> yyyyMMdd where:
>
> yyyy is the four-digit year
> MM is the two-digit month
> dd is the two-digit day

Discussions

What is a data steward?

Typically the entity tasked with maintaining and managing distribution of the dataset in question would be considered the steward of the data. Often this maintenance is not a GIS task alone, rather it is part of another business process to which the GIS tools have been applied successfully. In some instances the steward will not be the entity that actually creates the geographic data, but rather maintains an attribute that is attached to geography like county contacts for emergency response. The steward for the contacts has significantly altered the county dataset and becomes the steward of the resultant dataset. Additional discussion of stewardship roles and responsibilities are underway.

What is publishable data?

The meaning of "publishable data" is in flux. Most simply, publishable data is not sensitive in any respect and can be freely shared without reservation or restriction. However, there are instances when sensitive data must be shared with appropriate precautions on the part of both the data provider and data consumer. Work is in progress to further define those roles and necessary precautions. This naming convention recommendation is intended for use across the spectrum of meanings for publishable data.

What is a recommended best practice?

This naming convention is offered as a recommended best practice and does not carry with it requirements for compliance. A recommended best practice may become a requirement when needed. For instance, a naming convention could be a recommended best practice here in this arena but it could be used as a contract requirement for data that is to be delivered as part of that contract.

Is there a standard thesaurus for keywords?

Some users have asked whether there exist guidance on developing consistent keywords. There are several, each for specific communities like framework data, facilities management, environmental, telecommunications etc. though they may not necessarily speak to file naming specifically. We encourage the various communities of interest to contact the IGIC Standards and Recommendations committee so that we can post the various guidance documents in a central location for other data stewards to consider.